Qual è il contenuto del genoma?

Come interpretare una sequenza?

Localizzazione dei geni in una sequenza di DNA

- Esame della sequenza con il computer nel tentativo di identificare caratteristiche spesso associate a geni

Bioinformatica

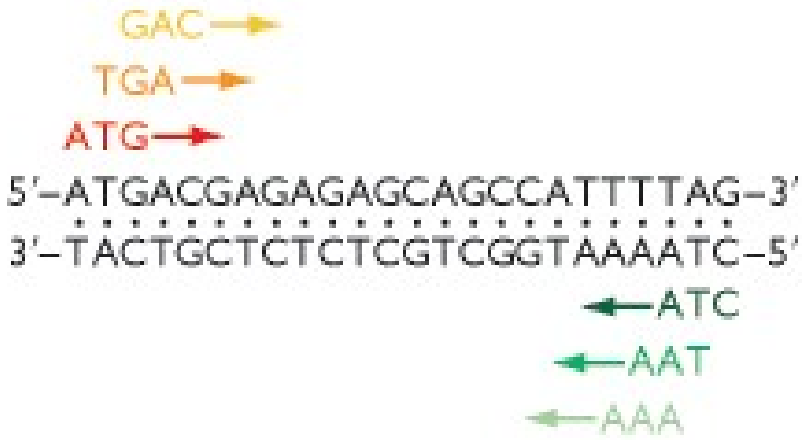
- Identificare un gene mediante **approcci sperimentali**

Dedurre dalla sequenza genomica i geni che codificano per le proteine

Individuazione dei moduli di lettura aperti (ORF) mediante analisi computazionale

- Ogni ORF presenta un codone di inizio (ATG) ed un codone di terminazione (TAA, TAG, TGA)

- Ogni sequenza di DNA ha 6 possibili cornici (schemi) di lettura diverse



Il punto chiave della corretta ricerca delle ORF sta nella frequenza con cui i codoni di terminazione appaiono in una sequenza di DNA

- Se il DNA ha una sequenza casuale di nucleotidi ed un contenuto in CG del 50%, le sequenze corrispondenti ai 3 codoni di terminazione si presenteranno con una frequenza di 4^3 (64nt).

  Se CG>50% i tre stop (ricchi in AT) avranno una frequenza minore (100-200nt ca.)

PERO'….

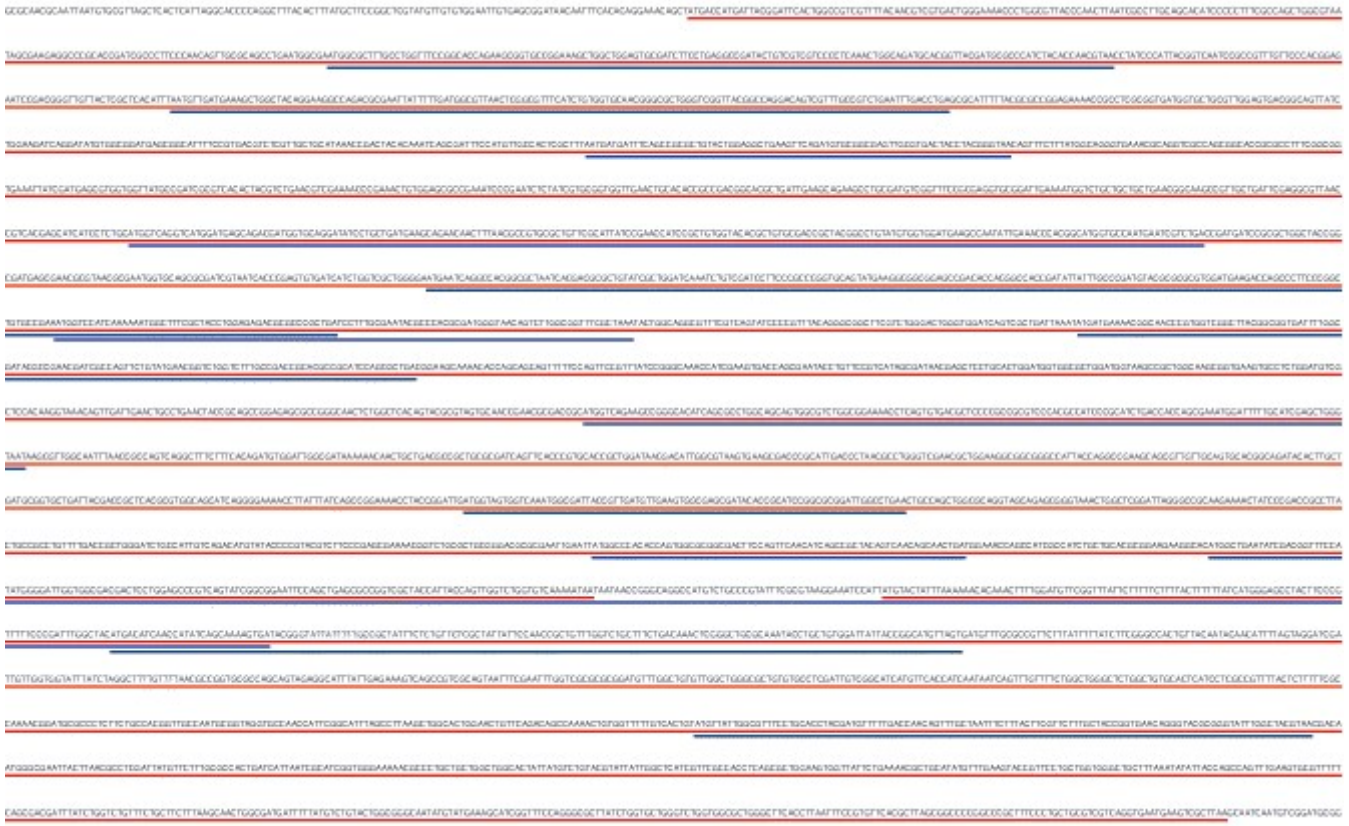
le ORF hanno una lunghezza decisamente maggiore (uomo 450 codoni).

QUINDI…

se una sequenza non presenta codoni di stop per almeno 300 nt (100 codoni) è molto probabilmente una vera ORF

La scansione di una sequenza alla ricerca di una ORF permette di localizzare i geni nel genoma batterico.

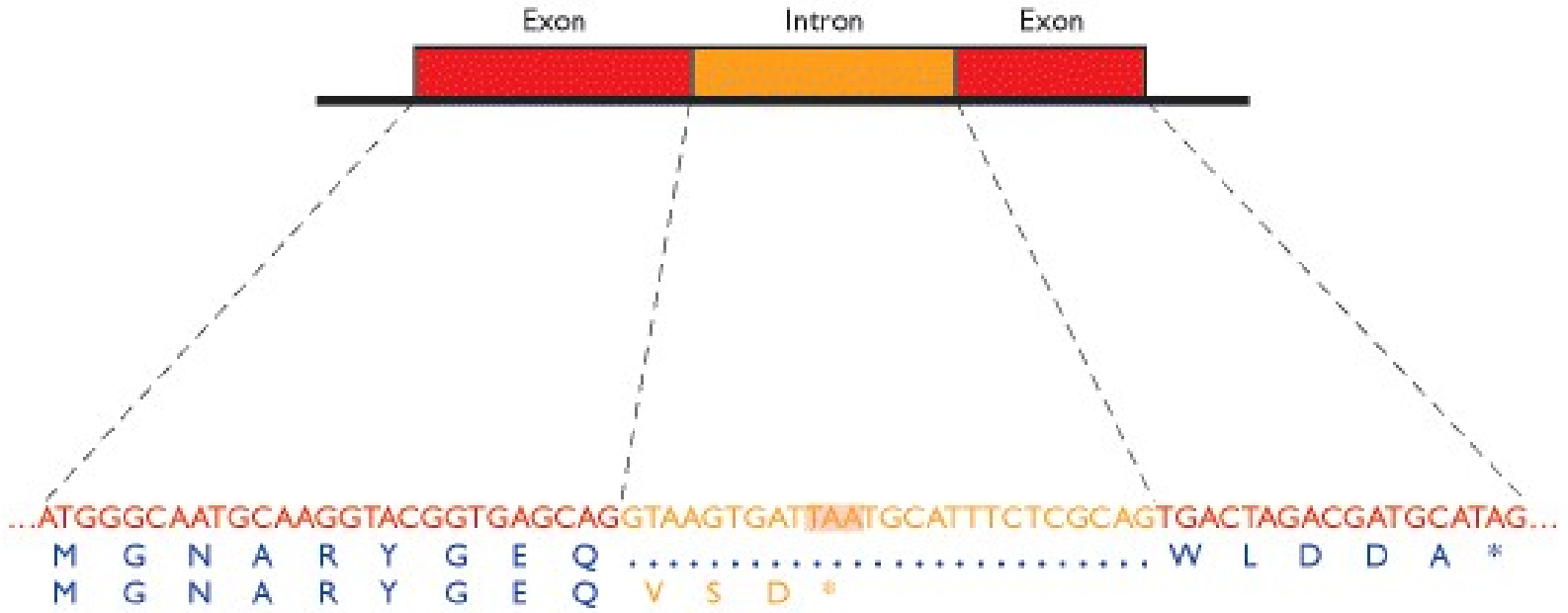
Il diagramma mostra 4522 bp dell'operone del lattosio di E. Coli in cui risultano sottolineate tutte le ORF di lunghezza superiore ai 50 codoni. La sequenza contiene due geni reali - *lacZ* e *lacY* – sottolineati in rosso. Questi geni sono facilmente ed inequivocabilmente identificati in quanto molto più lunghi delle ORF spurie sottolineate in blu.



L'analisi delle ORF è più problematica con il DNA eucariotico

- Nel DNA eucariotico c'è molto spazio fra i geni (il 70% del genoma umano è intergenico)

- Introni! I geni degli eucarioti sono discontinui (esoni spesso più corti di 100 codoni)



Modifiche alla procedura di analisi bioinformatica delle ORF introdotte per minimizzare il disturbo dovuto alla presenza degli introni

1) I software per l'analisi delle ORF prendono in considerazione i **codoni preferenziali.**

Preferenzialità (specie-specifica) nell'uso dei codoni (*codon bias*)

Es. leucina TTA, TTG, CTT, CTC, CTA, CTG ma nell'uomo in genere solo CTG; valina GTG più frequentemente di GTA

SI INSERISCONO NEI SOFTWARE PER L'ANALISI DELLE ORF I CODONI PREFERENZIALI DEGLI ORGANISMI DA STUDIARE (I QUALI APPARIRANNO NELLE ORF Più FREQUENTEMENTE DI QUANTO LA CASUALITA' IMPONGA E DI QUANTO APPARIRANNO I SINONIMI)

2) ricerca delle **giunzioni di splicing**

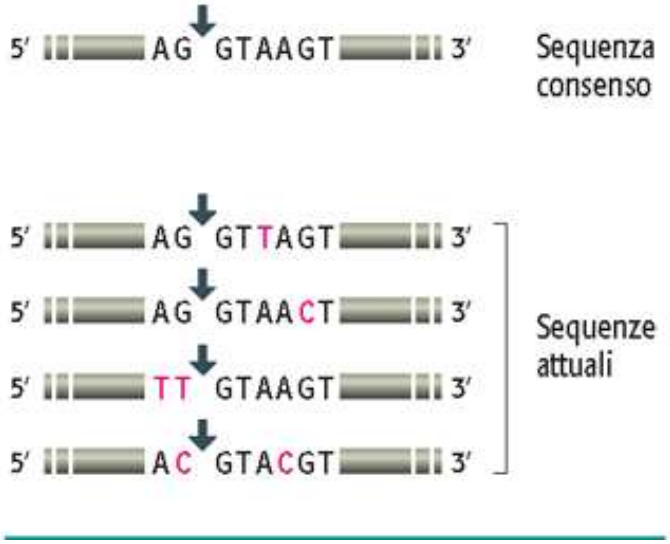
esone-introne

5′ AG|GTAAGT 3′
Consensus

introne-esone

5′ PyPyPyPyPyPyN CAG | 3
Consensus

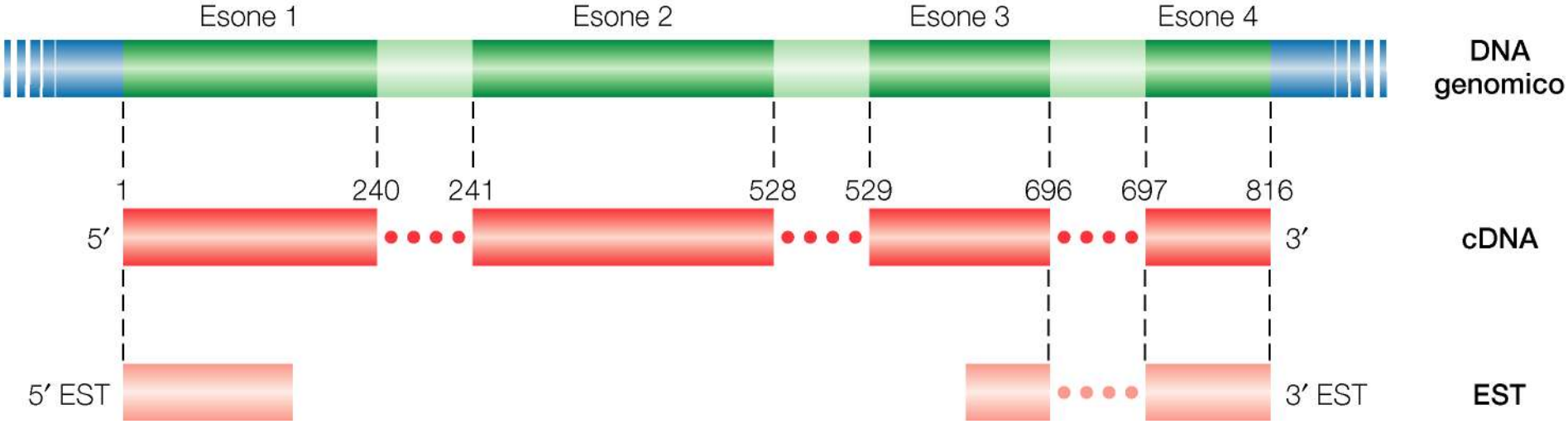
rappresentano "consensus" con pochi nt veramente invariabili



Nelle giunzioni esone-introne al 5′ soltanto la sequenza "GT" che segue immediatamente il sito di splicing è conservata!!!

4) Le ORF possono essere pescate da collezioni di cDNA o di EST (evidenza diretta della trascrizione di una sequenza)

EST: Espressed Sequence Tags, parte di geni espressi che possono essere utilizzati come sonde, ad es. nei microArray oppure nella scoperta di nuovi geni, nella determinazione della loro sequenza e della loro posizione nel genoma.



Allineamento con il DNA genomico di un cDNA completamente sequenziato e di sequenze EST

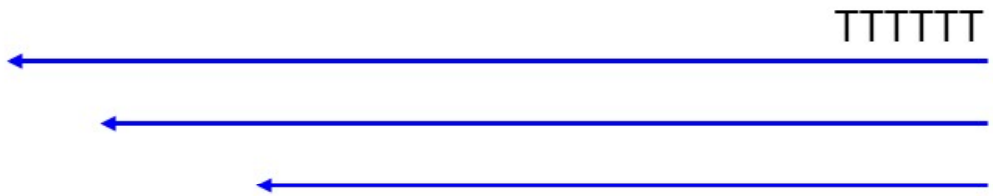
IDENTIFICARE LA SEQUENZA DELL'UNITA' DI TRASCRIZIONE

SPLICING

LOCALIZZARE L'ORF

La costruzione del cDNA

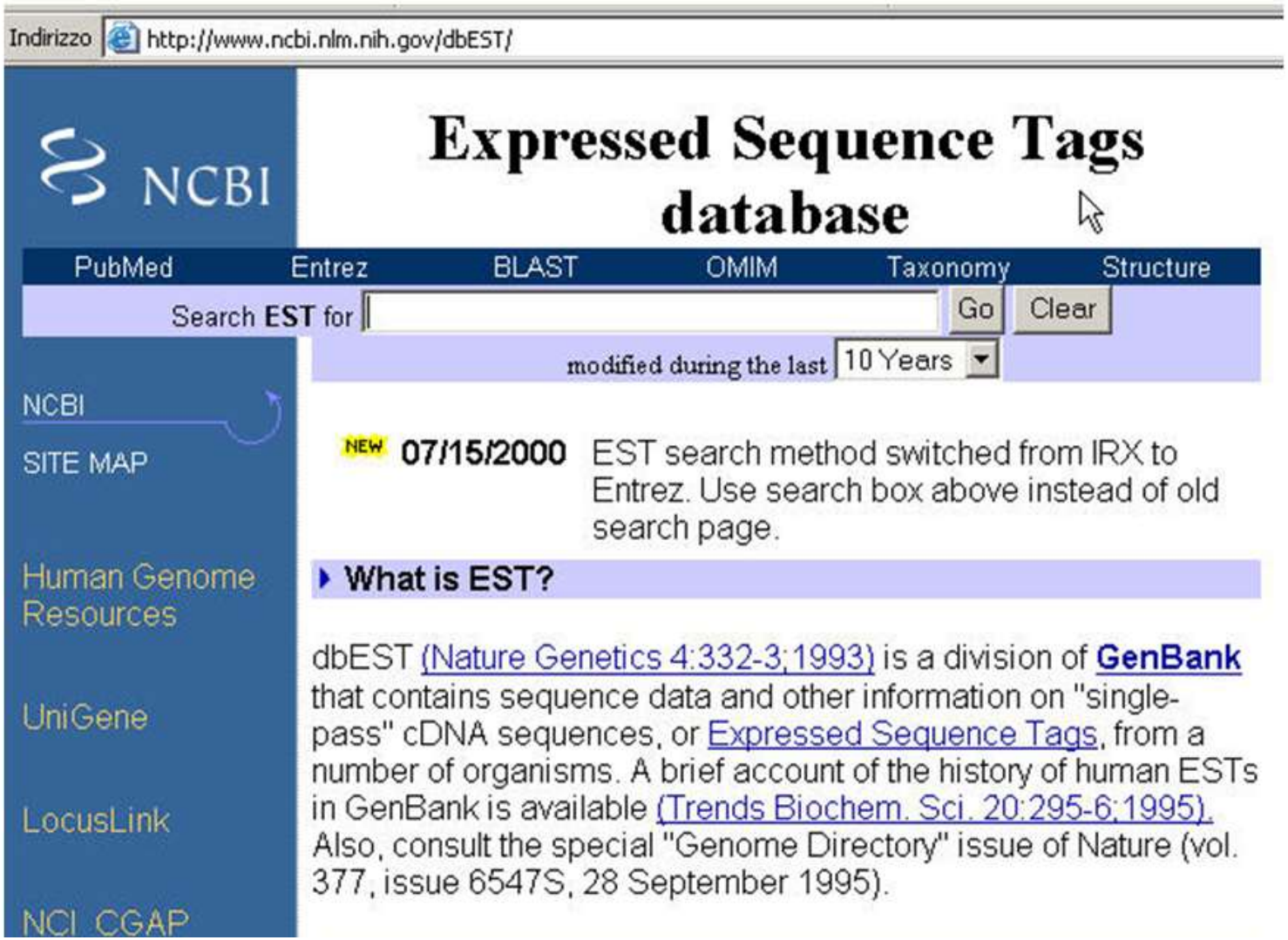
Le sequenze di cDNA ottenute dall'mRNA sono generalmente tronche



cDNA, EST e banche dati

dbEST (pronuncia 'the best')

Divisione di GenBank che contiene tutte le sequenze EST, classificate per specie, tessuto, patologia…

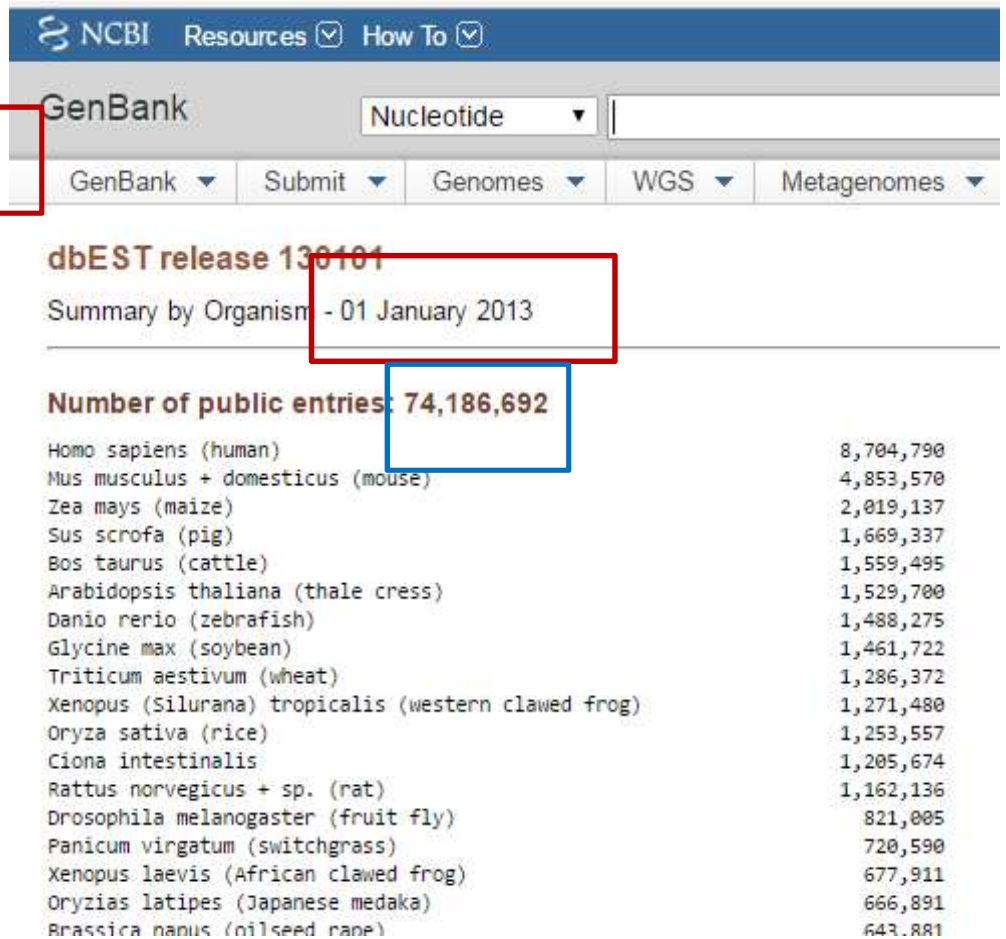


dbEST release 103103

Summary by Organism - October 31, 2003

Number of public entries: 18,971,362

Homo sapiens (human)	5,427,521
Mus musculus + domesticus (mouse)	3,915,334
Rattus sp. (rat)	538,251
Triticum aestivum (wheat)	500,902
Ciona intestinalis	492,488
Gallus gallus (chicken)	451,565
Zea mays (maize)	383,759
Danio rerio (zebrafish)	362,445
Hordeum vulgare + subsp. vulgare (barley)	348,233
Xenopus laevis (African clawed frog)	344,747
Glycine max (soybean)	341,578
Bos taurus (cattle)	329,387
Drosophila melanogaster (fruit fly)	261,414
Oryza sativa (rice)	260,890
Saccharum officinarum	246,301
Caenorhabditis elegans (nematode)	215,200
Silurana tropicalis	209,240
Arabidopsis thaliana (thale cress)	190,732
Medicago truncatula (barrel medic)	187,763
Sus scrofa (pig)	171,920

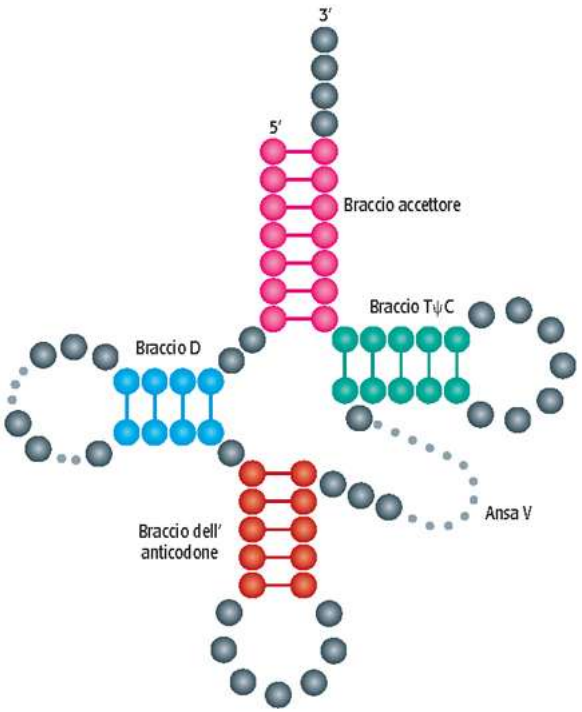


Individuazione di geni che codificano per RNA funzionali

Questi geni non contengono ORF.

Possibilità di APPAIAMENTO INTRAMOLECOLARE

(A) Struttura a quadrifoglio del tRNA



Tutti i tRNA si ripiegano in una caratteristica struttura a quadrifoglio che è stabilizzata da appaiamenti intramolecolari in quattro regioni diverse.

L'individuazione di queste regioni di complementarietà permette di individuare con una certa facilità i geni per i tRNA.

(B) Sequenza di uno dei geni di *Escherichia coli* $tRNA^{leu}$

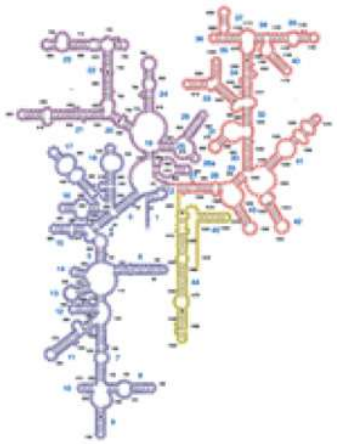
5′ GCCGAAGTGGCGAAATCGGTAGTCGCAGTTGATTCAAAATCAACCGTAGAAATACGTGCCGGTTCGAGTCCGGCCTTCGGCACCA 3′

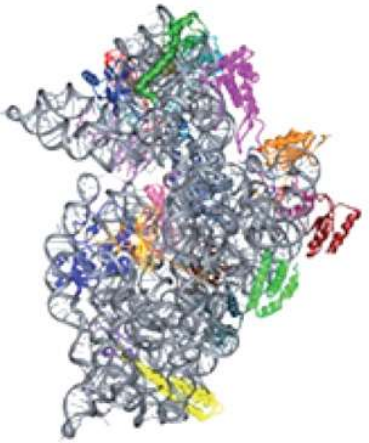
Anche gli rRNA e alcuni piccoli RNA funzionali adottano strutture secondarie con una complessità sufficiente da permettere di identificare i geni corrispondenti nel genoma.

A) struttura secondaria dell'rRNA

B) struttura tridimensionale dell'rRNA

C) struttura della subunità ribosomale

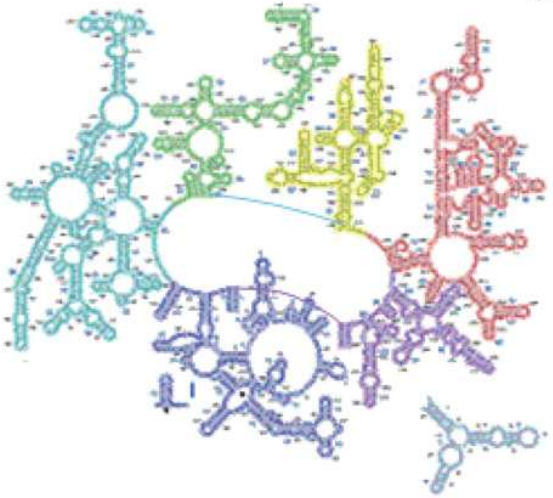


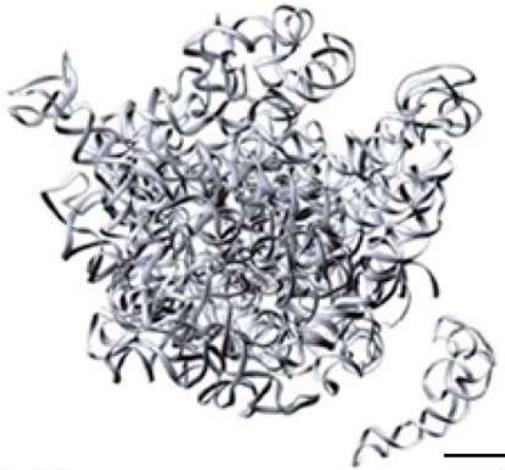


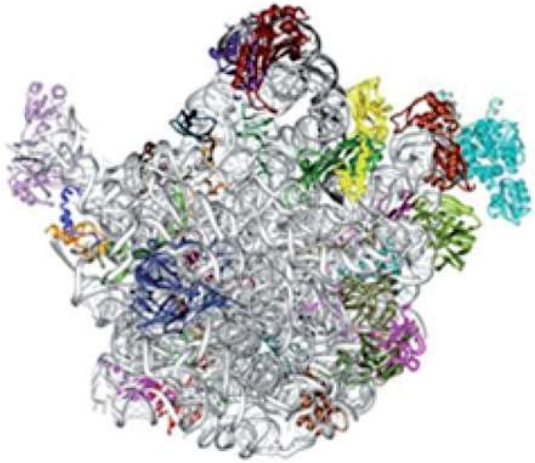
rRNA 16S

+21 r-proteine

subunità 30S



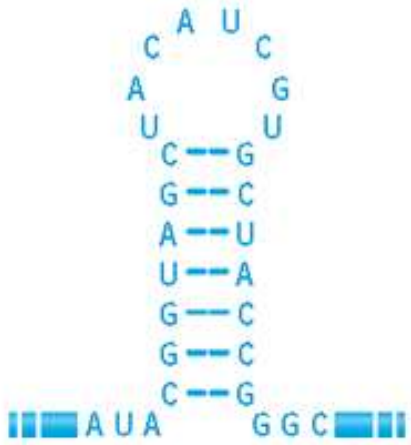




rRNA 23S+5S

+34 r-proteine

subunità 50S



Struttura a stem-loop di una molecola di RNA

Anche gli RNA funzionali che non assumono strutture secondarie complesse sono comunque caratterizzati da più o meno sequenze in grado di creare strutture a forcina.

Una sequenza compatibile con una forcina abbastanza stabile (un minimo di contenuto in CG) rappresenta un indicatore della presenza di un gene per RNA

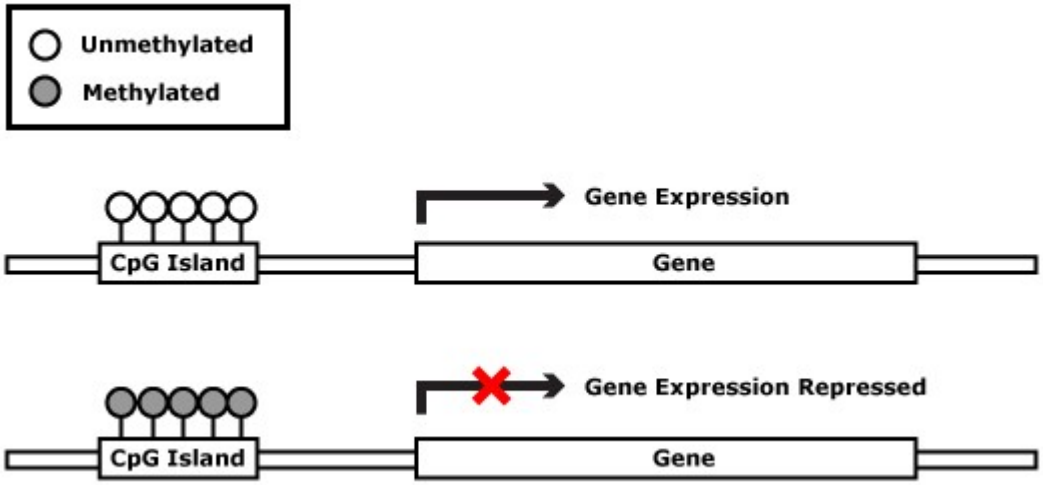
Anche nel caso di geni per RNA funzionali si possono individuare **sequenze regolative**, che sono diverse da quelle dei geni codificanti proteine. Alcune possono trovarsi anche all'interno del gene.

Nei genomi compatti bisogna fare molta attenzione al DNA che rimane dopo una estensiva ricerca e individuazione di geni codificanti proteine: **gli spazi "vuoti"** sono spesso occupati da geni per RNA

Regioni regolative al 5' dei geni

-Siti di riconoscimento di fattori trascrizionali

-Isole CpG (nei vertebrati e nell'uomo per ca. il 50% dei geni)*



*In una sequenza di DNA di un vertebrato queste isole sono un forte indizio che un gene inizi nella regione subito a valle.

Isole CpG

Identificare un'isola CpG significa con ogni probabilità incontrare poco più a valle un gene.

Le isole CpG, infatti, sono elementi di controllo epigenetico dell'espressione genica

Le isole CpG controllano l'espressione genica dei geni a valle sulla base del loro stato di metilazione

http://www.bioinformatics.org/sms2/cpg_islands.html

Isole CpG

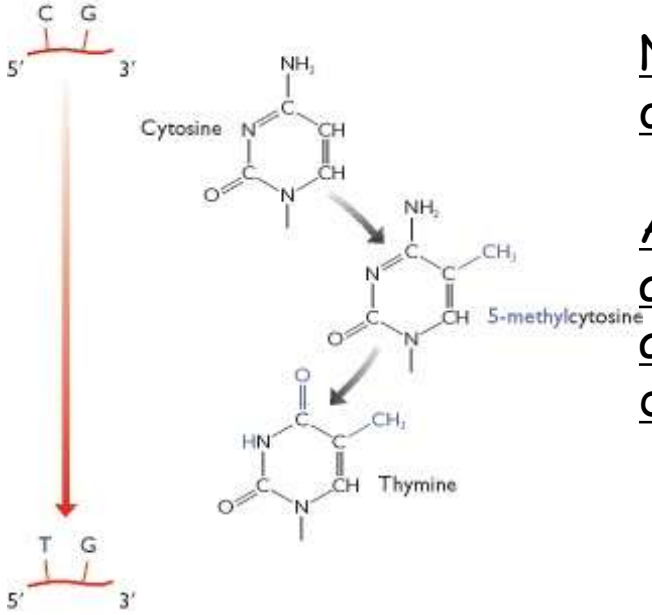
La metilazione di citosine è una delle più comuni modificazioni epigenetiche osservate nei genomi eucariotici. Nei vertebrati e nelle piante risulta metilato rispettivamente il 10% e il 30% delle citosine.

Le citosine metilate sono generalmente quelle presenti nel dinucleotide 5'-CpG-3'.

MA

La metilazione delle Citosine non è una cosa buona per il genoma

Isole CpG



La 5-metil-citosina è soggetta a **deaminazione** formando timina. Il risultato di questo processo è che il dinucleotide CpG è generalmente evitato nel genoma dei vertebrati e delle piante.

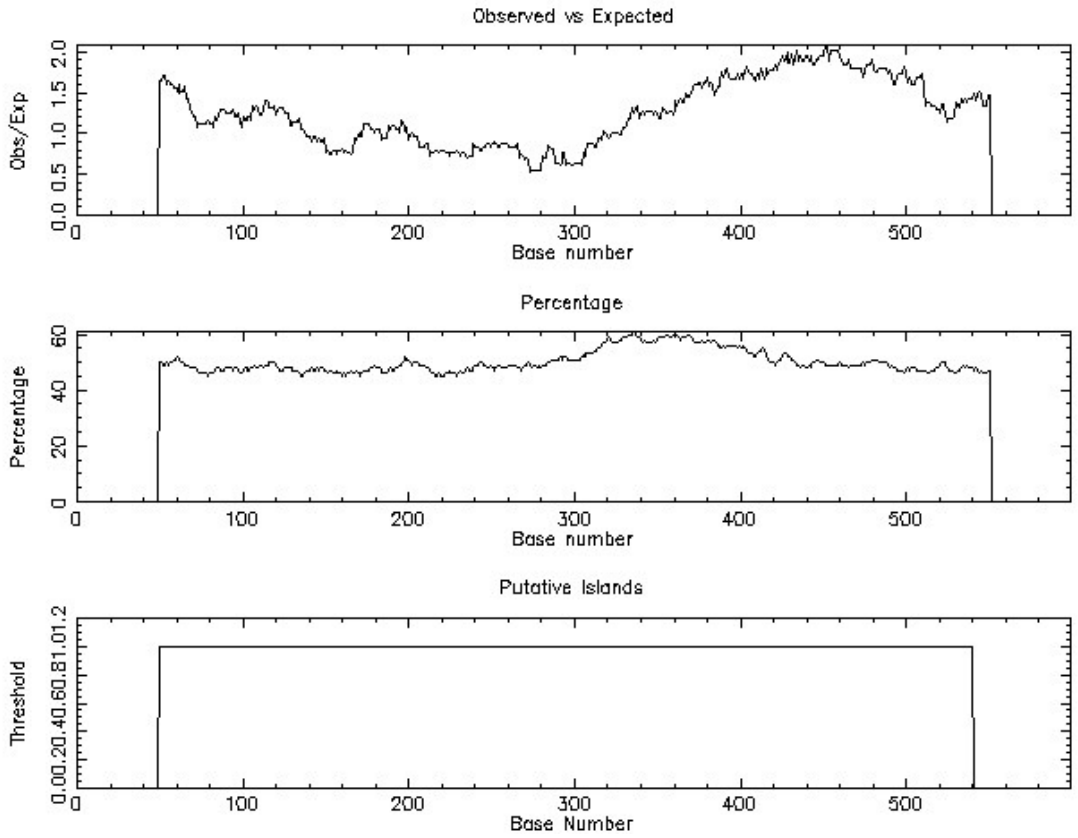
Nel genoma umano infatti la frequenza osservata di CpG è circa 1/5 di quella attesa.

Anche nelle isole, dove la frequenza è più alta, essa è comunque inferiore a quella attesa sulla base di una distribuzione casuale dei nucleotidi (circa 1/3)

Isole CpG

Nota la sequenza genomica è possibile predire la localizzazione delle isole GpG con programmi bioinformatici. La definizione operativa che viene comunemente utilizzata per la definizione di un'isola CpG nei mammiferi è la seguente:

- L > 500 bp
- C+G% > 55%
- CpG Obs/Exp* > 0,65

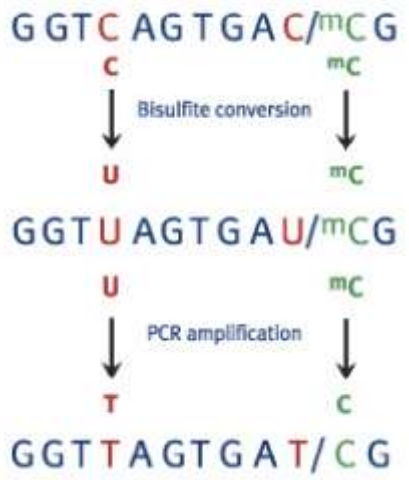


Isole CpG

Le isole CpG sono localizzate nella regione del promotore di circa il 50% dei geni umani, la maggior parte dei quali di tipo costitutivo (housekeeping, espressi in molti tessuti diversi = isole CpG non metilate con RNApol saldamente legata).

I geni con isole CpG metilate sono generalmente tessuto specifici (si esprimono dove le loro isole non sono metilate).

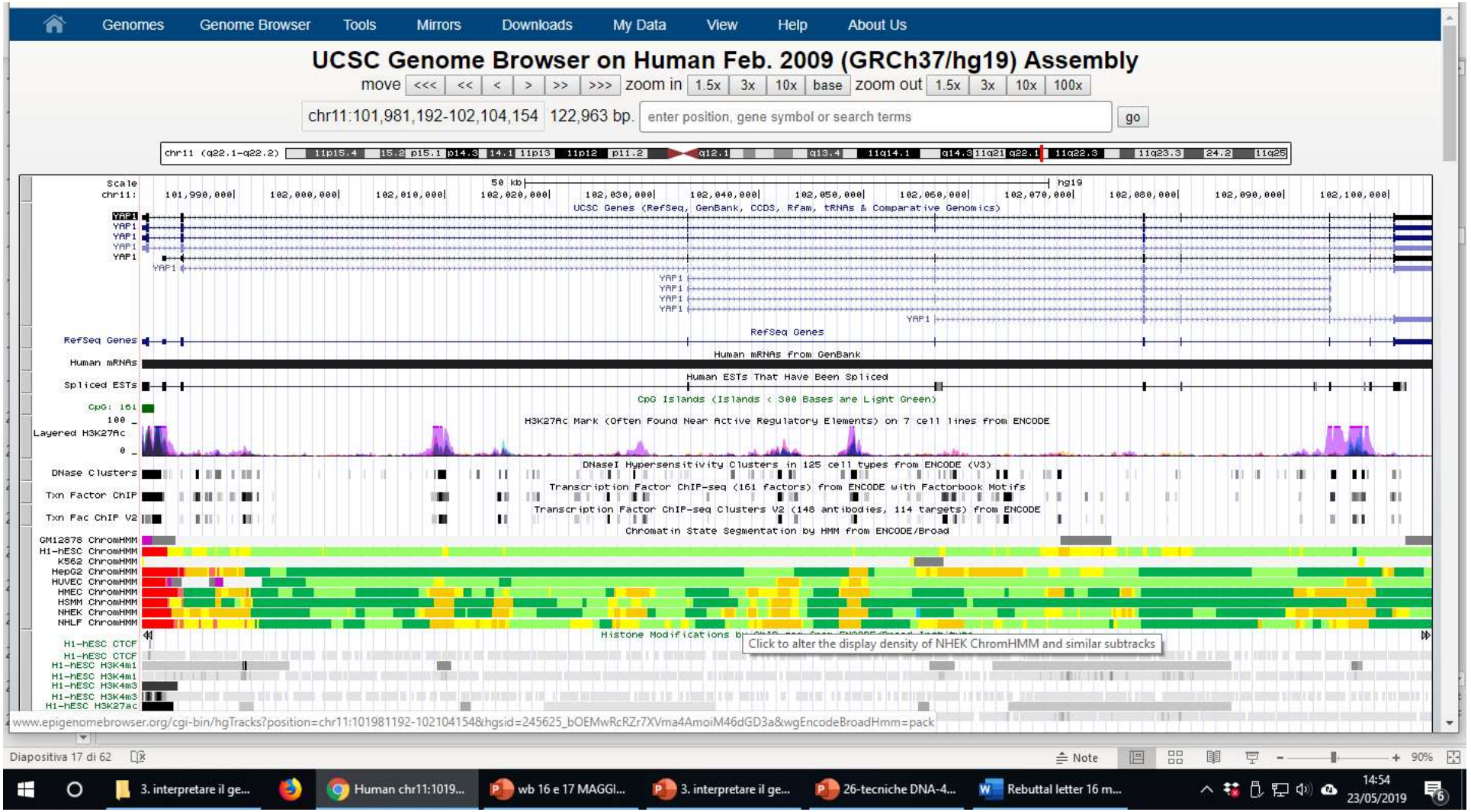
Il **trattamento del DNA con bisolfito di sodio** induce la modifica (attraverso una serie di intermedi) della citosina non metilata in uracile

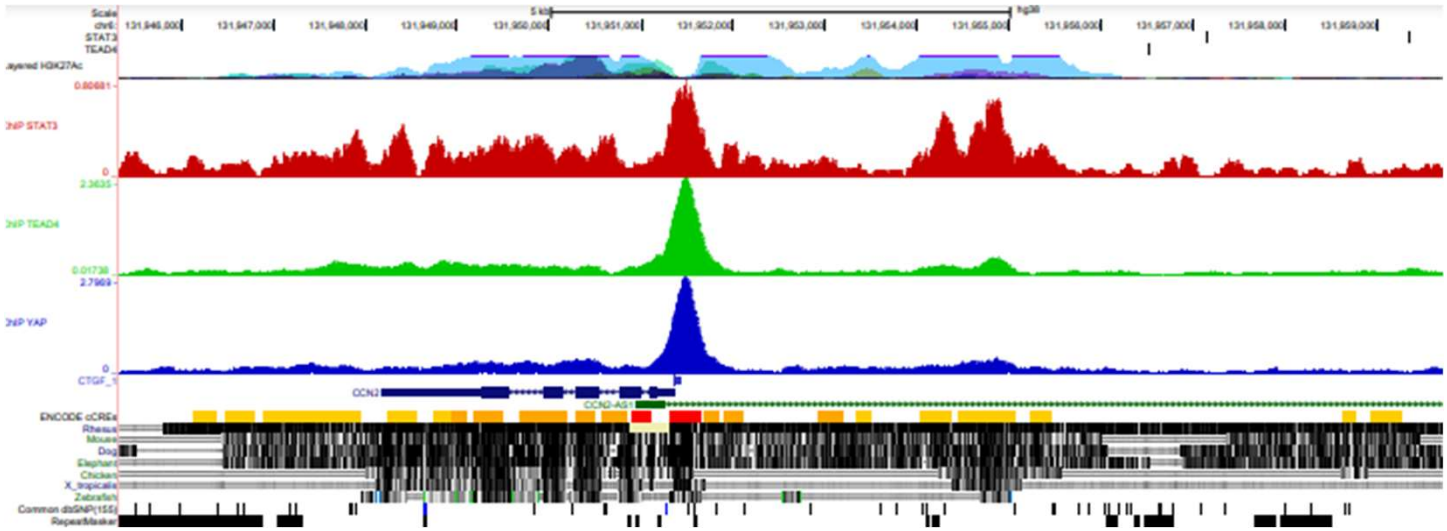


La PCR amplifica il DNA modificato introducendo T al posto delle C

Lo stato di metilazione di un'isola CpG quindi si valuta sequenziando lo stesso tratto di DNA con o senza bisolfito: le C che rimangono tali sono quelle metilate.

La ChIP seq analysis ha permesso di mappare il genoma con siti per fattori trascrizionali e con modifiche epigenetiche




Scale
chr6
131,946,000
131,947,000
131,948,000
131,949,000
131,950,000
131,951,000
131,952,000
131,953,000
131,954,000
131,955,000
131,956,000
131,957,000
131,958,000
131,959,000
5 kb
hg38
STAT3
TEAD4
ayered H3K27Ac
0.00681
ChIP STAT3
0
2.3635
ChIP TEAD4
0.01738
2.7969
ChIP YAP
0
CTGF_v
CCN2
CCN2-AS1
ENCODE cCREs
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Common dbSNP(155)
RepeatMasker

Le analisi di omologie di sequenza sono un ulteriore strumento per l'identificazione dei geni

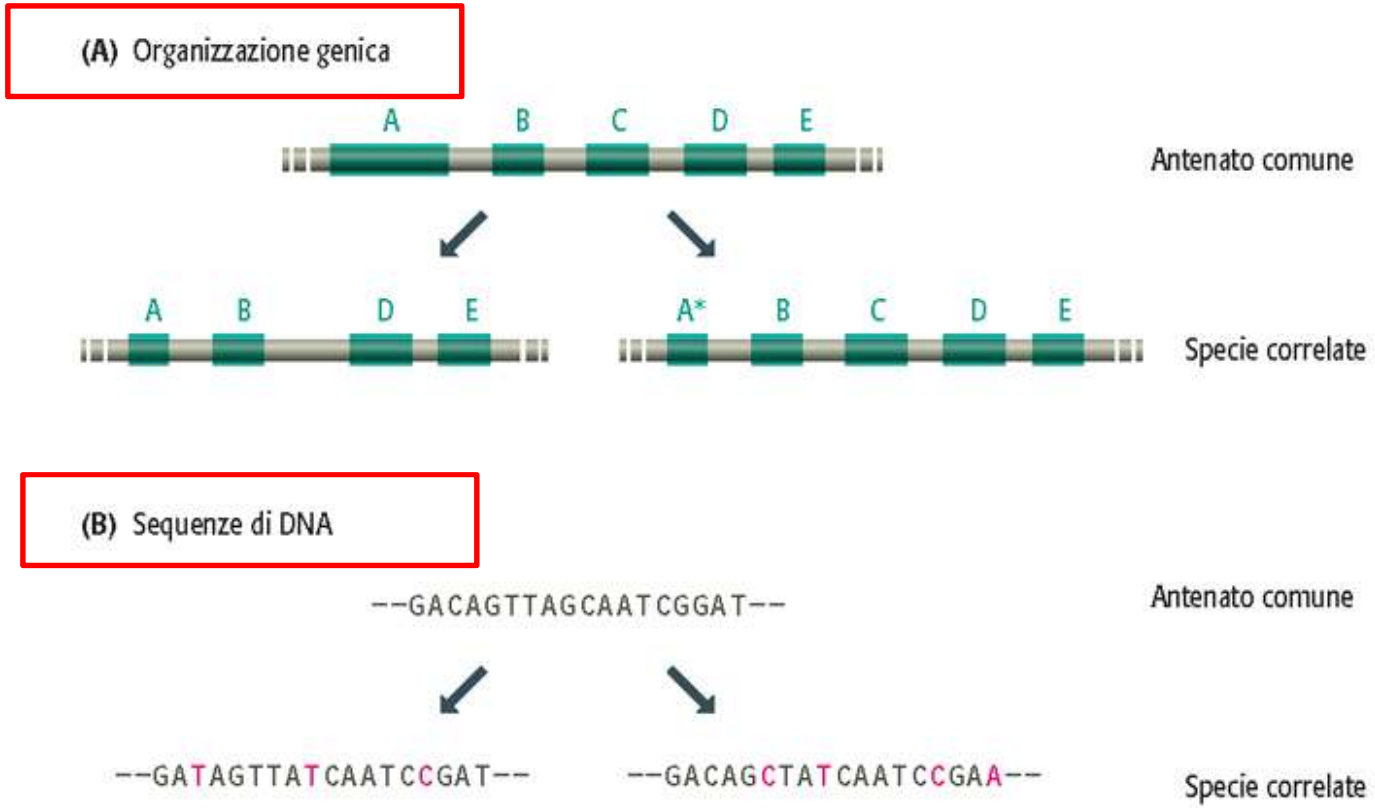
Confronto tra la sequenza in esame e tutte le sequenze presenti in banca dati, cercando similarità o identità con geni già sequenziati.

L'omologia può indicare geni correlati evolutivamente.

L'analisi, oltre a permettere la validazione di esoni della cui ORF non si è certi, serve anche ad assegnare funzioni ad un gene appena scoperto

Genomica comparata

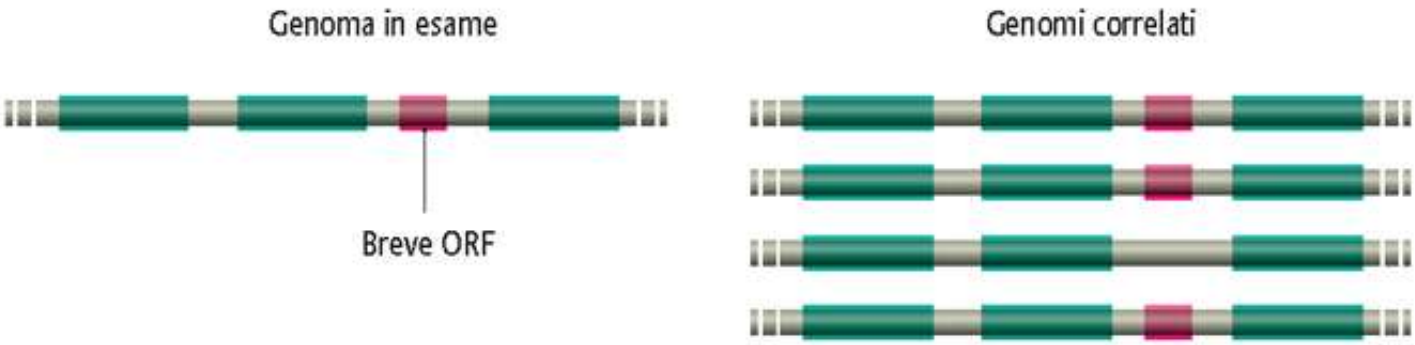
Quando si confrontano genomi di specie correlate, i geni omologhi sono facilmente identificabili poiché hanno un elevato grado di similarità. I **geni omologhi** rappresentano geni correlati evolutivamente.



La comparazione completa di genomi di **specie correlate** rivela come **le somiglianze risiedano di più all'interno dei geni** che non nelle sequenze intergeniche.

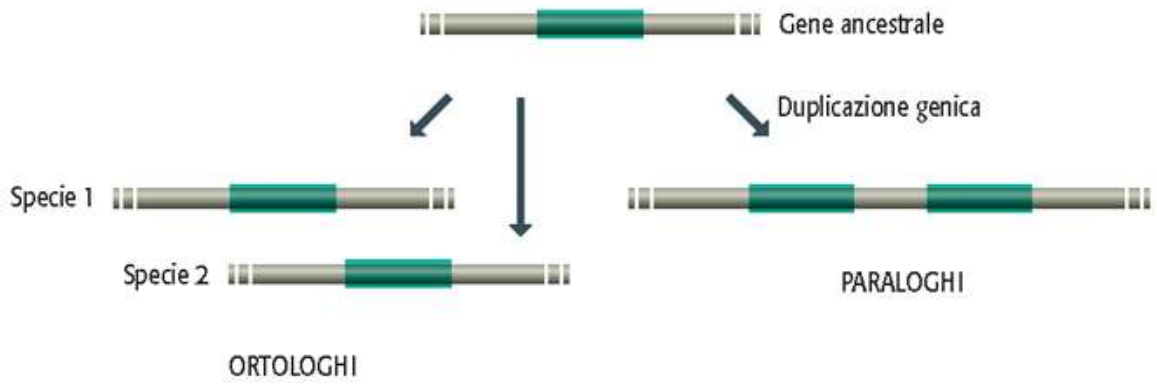
Quando si confrontano **genomi correlati** i geni omologhi sono facilmente identificabili e qualsiasi ORF che non ha un chiaro omologo nel secondo genoma può essere tranquillamente scartata in quanto sequenza casuale.

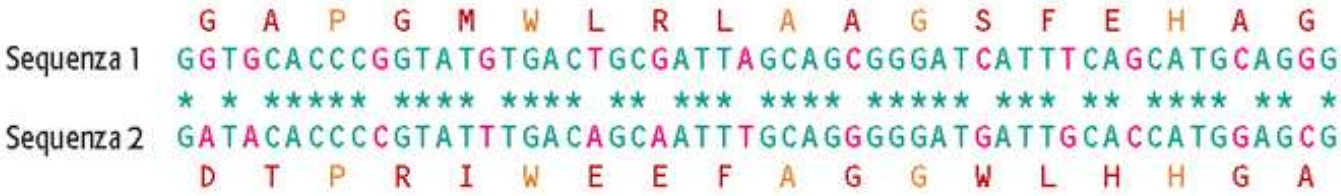
La **sintenia**, cioè la conservazione dell'ordine dei geni, rende ancora più efficace la genomica comparata.



In questo esempio l'ORF in esame è presente in tre su quattro dei genomi correlati per cui è probabile che si tratti di una ORF vera

Oltre ai geni **ORTOLOGHI**, cioè geni presenti in specie diverse il cui antenato comune esisteva prima della divisione delle due specie, ci sono i geni **PARALOGHI**, cioè geni omologhi presenti nello stesso organismo (e non necessariamente presenti in specie antenate) che si sono formati per duplicazione di un gene ancestrale.





Non confondere i concetti di omologia e similarità. Solo quest'ultima è quantificabile.

I geni omologhi posso essere organizzati in **famiglie geniche**

I membri di una stessa famiglia genica possono essere localizzati in unico cluster, dispersi, o localizzati in più cluster:

Geni in cluster:

α-globin (7), growth hormone (5), Class I HLA heavy chain (20),….

Geni dispersi:

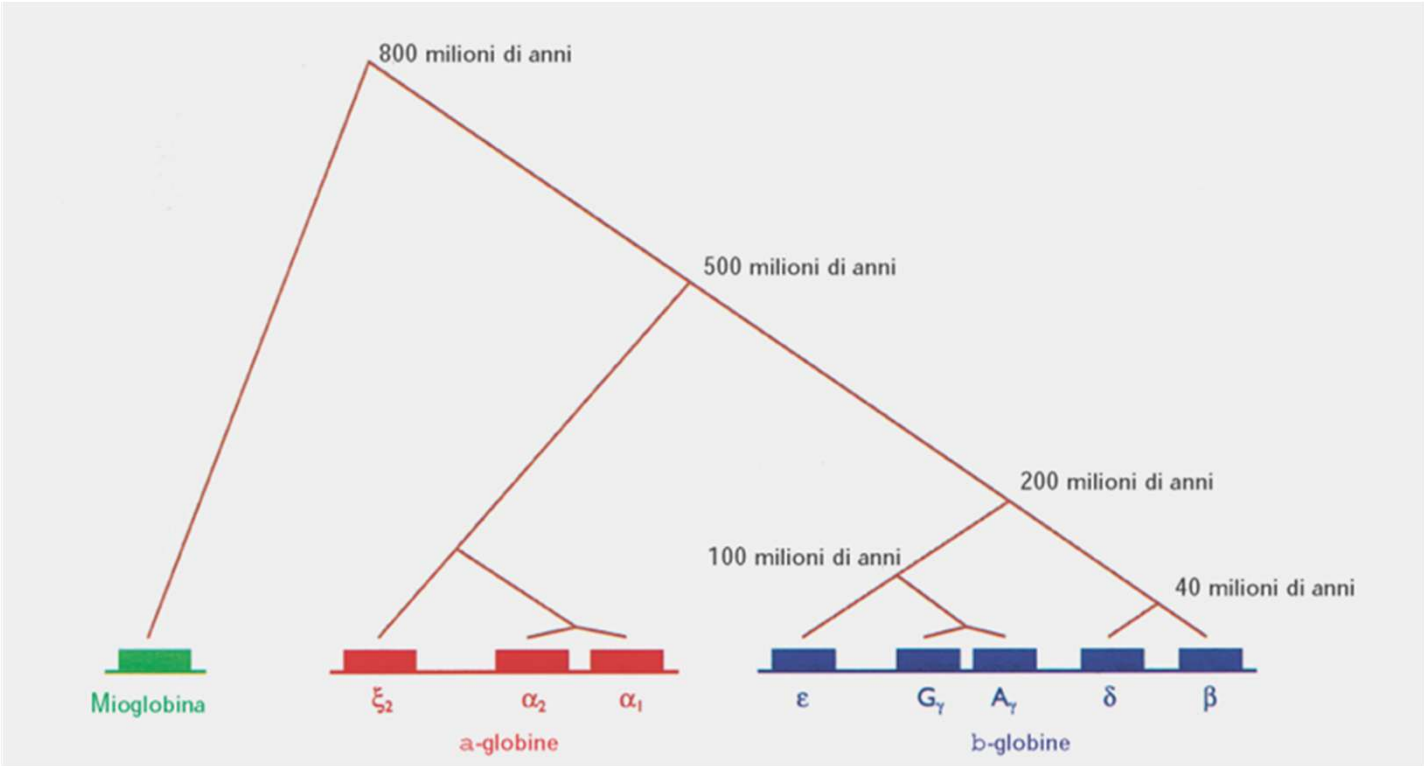
Pyruvate dehydrogenase (2), Aldolase (5), PAX (>12),..

Geni localizzati in più cluster:

HOX (38 – 4), Histones (61 – 2), Olfactory receptors (>900 – 25),…

La duplicazione genica è uno dei meccanismi attraverso cui si sono formate le **famiglie geniche**

Duplicazione genica nell'evoluzione della famiglia genica delle globuline umane



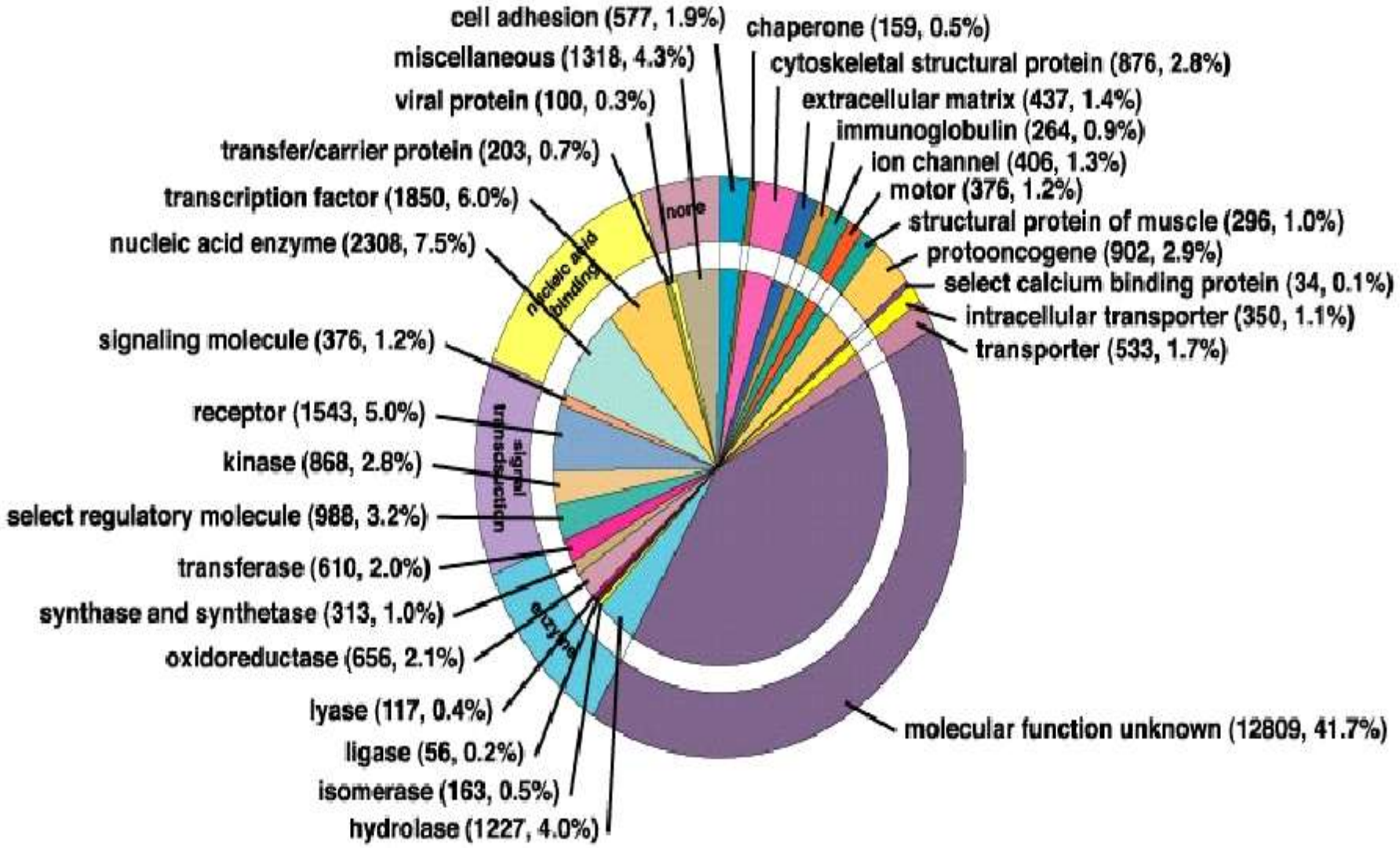
Assegnazione sperimentale della funzione di un gene

- Oggi possiamo assegnare funzioni (fenotipo) al numero enorme di geni identificati grazie ai diversi Progetti Genoma

- I risultati ottenuti mediante l'analisi di omologia non ci permettono di assegnare una funzione a tutti i geni; essi devono essere naturalmente integrati e confermati da studi sperimentali.

La funzione dei geni eucariotici

La funzione di una grossa frazione dei geni umani rimane sconosciuta



Analisi genetica classica:

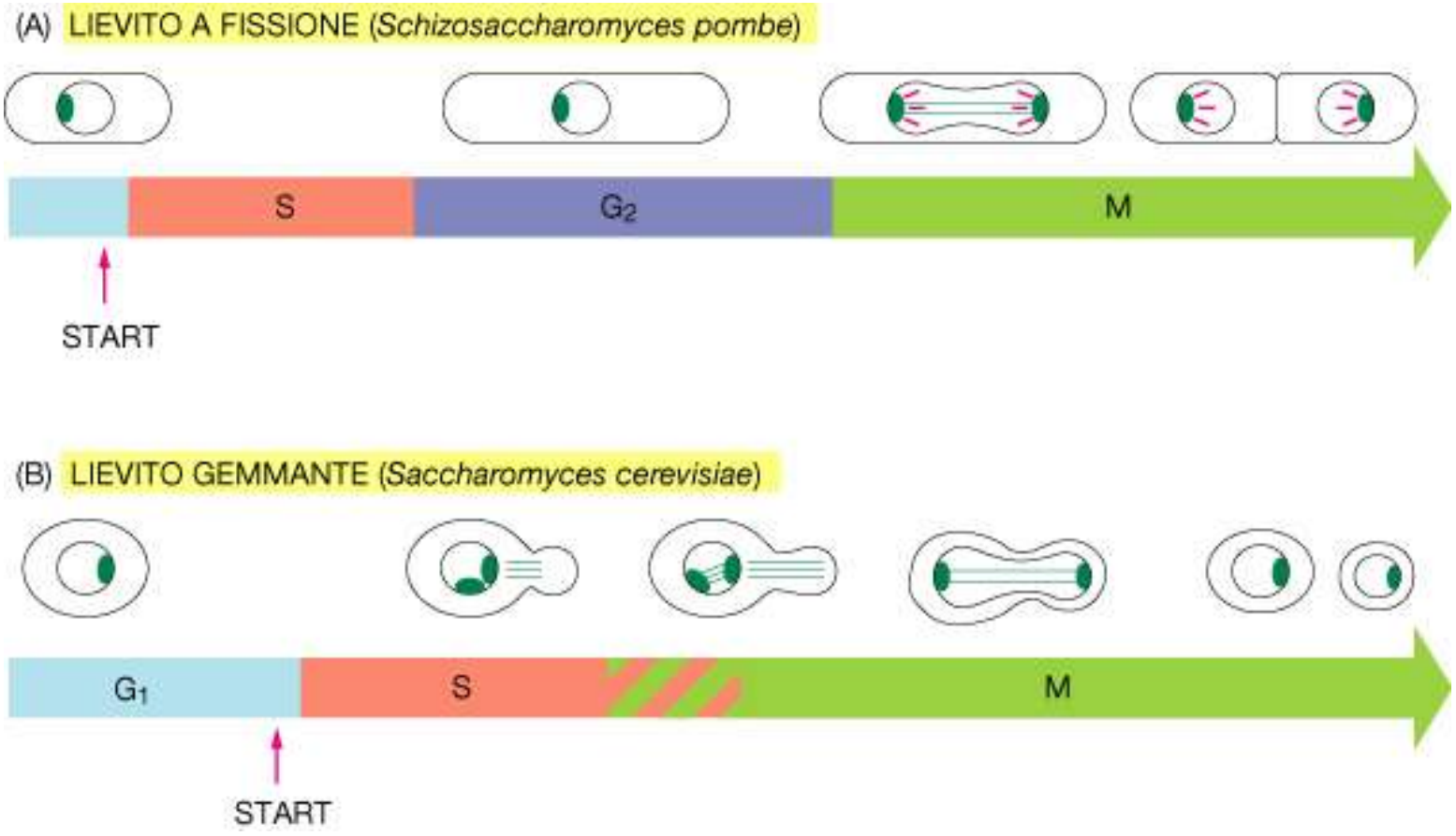
cerchiamo le basi genetiche di un fenotipo analizzandone i mutanti (indotti sperimentalmente o presenti in natura)

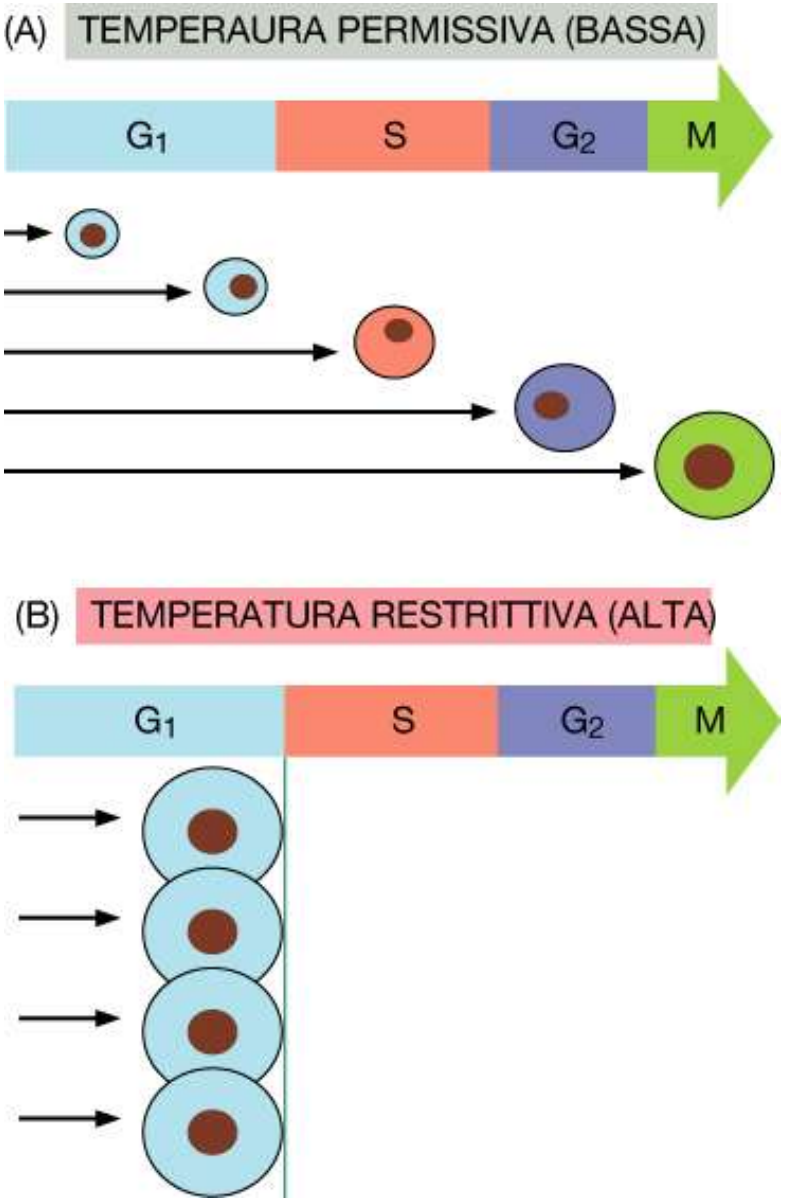
Es. identicazione di geni che controllano determinate funzioni cellulari:

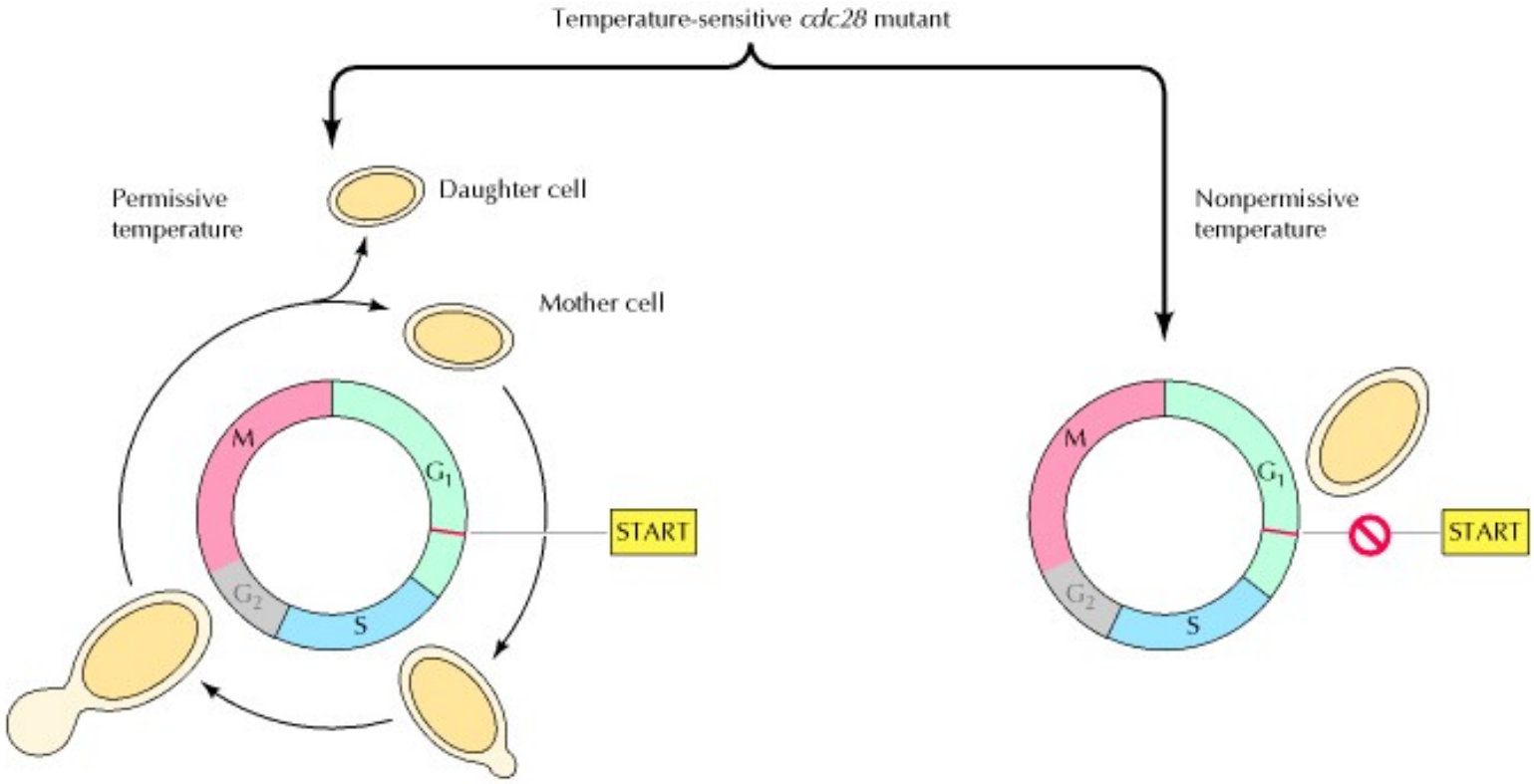
Es. 1 Lieviti Mutanti Temperatura Sensibili del ciclo cellulare

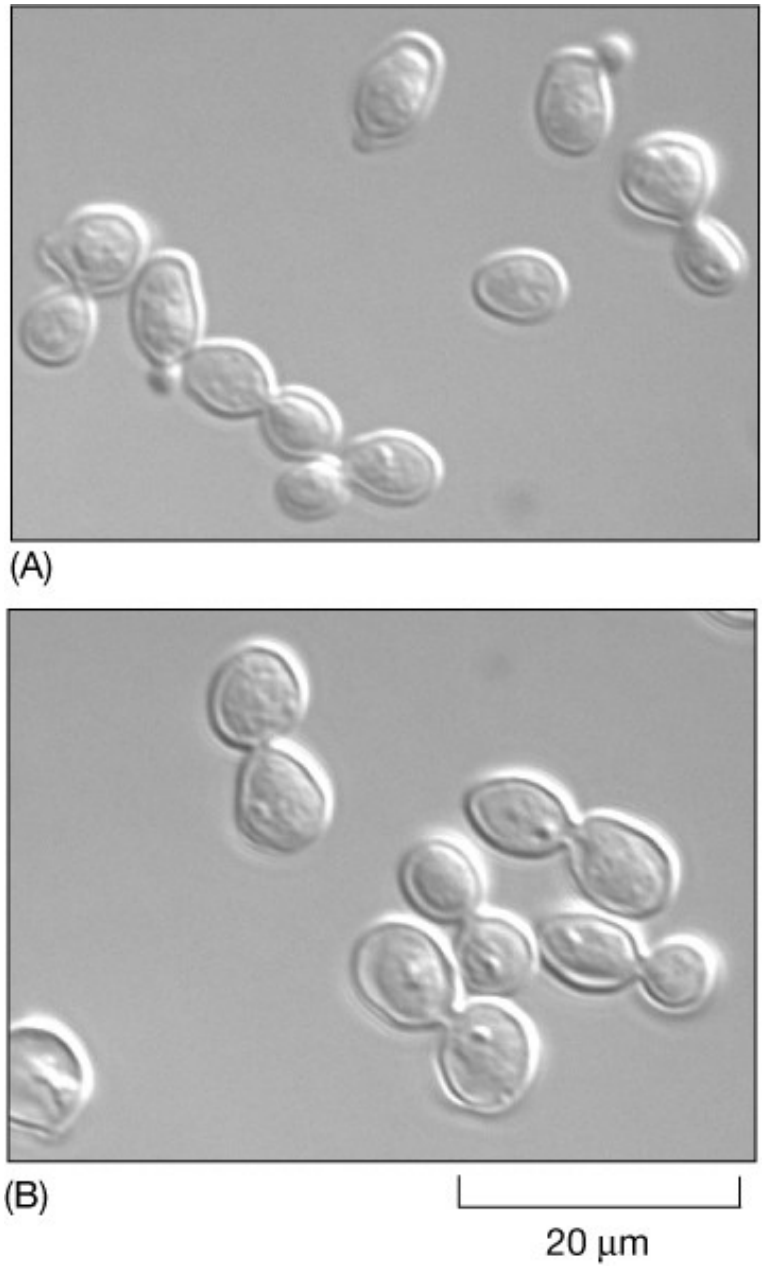
Es. 2 *C. Elegans* mutanti dell'apoptosi

Es. 3 Topo Shaker 2 -/- mutanti della funzione uditiva


(A) LIEVITO A FISSIONE (Schizosaccharomyces pombe)
S
G2
M
START
(B) LIEVITO GEMMANTE (Saccharomyces cerevisiae)
G1
S
M
START


(A) TEMPERAURA PERMISSIVA (BASSA)
G1
S
G2
M
(B) TEMPERATURA RESTRITTIVA (ALTA)
G1
S
G2
M


Temperature-sensitive *cdc28* mutant
Permissive temperature
Daughter cell
Mother cell
Nonpermissive temperature
M
G_1
START
G_2
S
M
G_1
START
G_2
S


(A)
(B)
20 µm

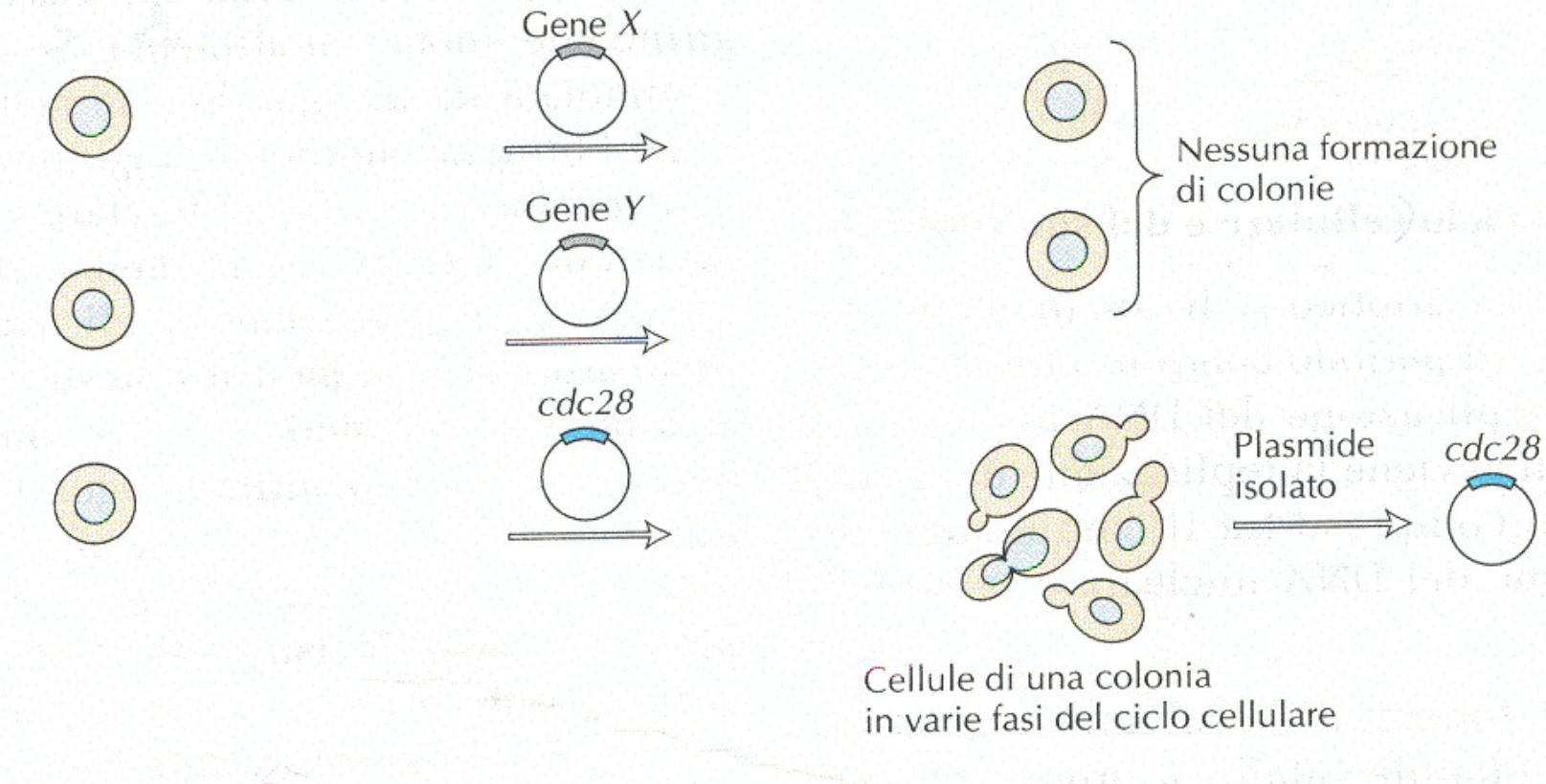
Cellule mutate a 25°C (temperatura permissiva)

Trasfezione con libreria plasmidica di DNA di lievito normale

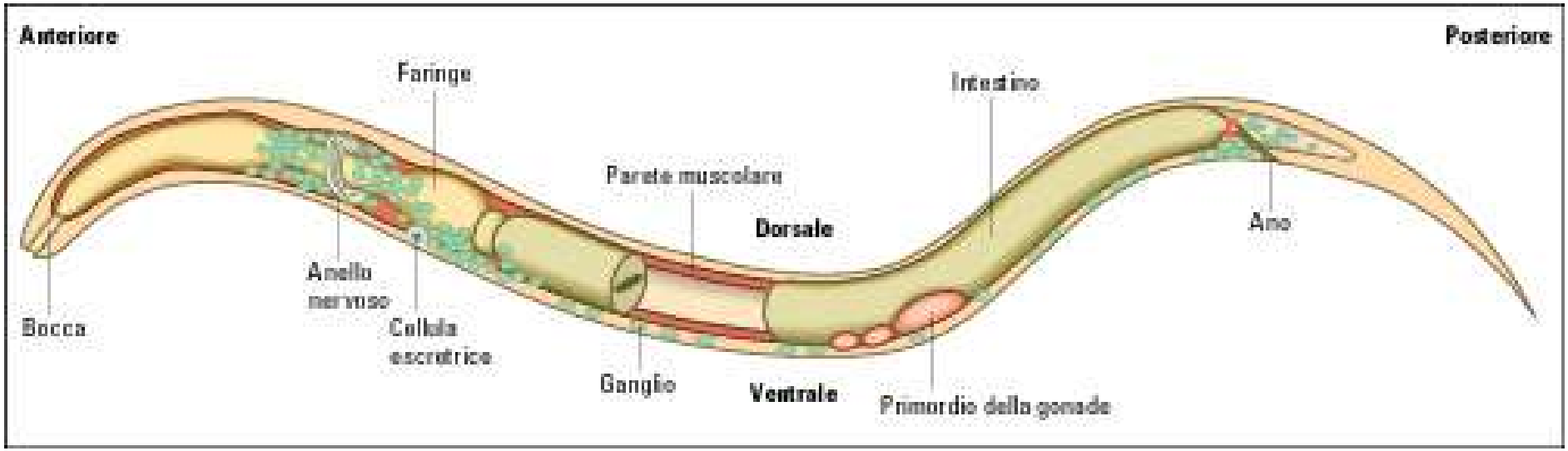
Semina dei lieviti trasfettati e coltura a 35°C (temperatura non permissiva)

Isolamento dei cloni che hanno complementato la mutazione

Isolamento del plasmide e quindi del gene cdc normale


Le cellule *cdc28*^{ts} a 25 °C
Transfezione con libreria plasmidica di DNA di tipo selvatico di *S. cerevisiae*
Gene *X*
Gene *Y*
cdc28
Le cellule transfettate *cdc28*^{ts} crescono a 35 °C
Nessuna formazione di colonie
Plasmide isolato
cdc28
Cellule di una colonia in varie fasi del ciclo cellulare

Caenorhabditis elegans as a perfect model for programmed cell death



1090 cells

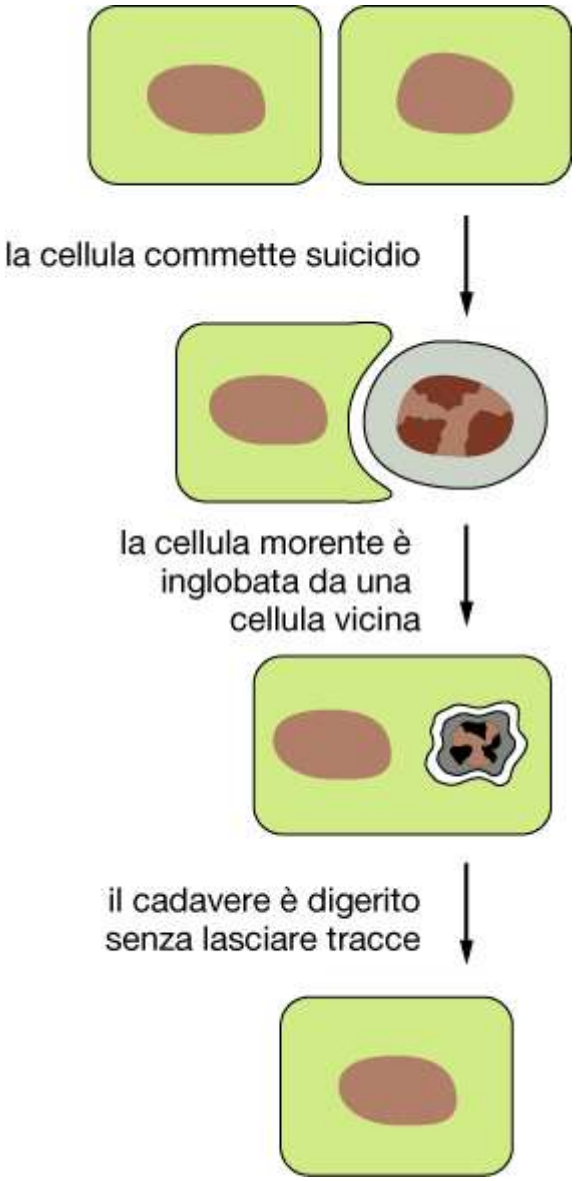
131 die

959 survive

In mutanti dei geni Ced-3 o Ced-4 le 1090 cellule sopravvivono tutte (Geni pro-apoptotici).

In mutanti del gene Ced-9 le 1090 cellule muoiono tutte (gene anti-apoptotico).

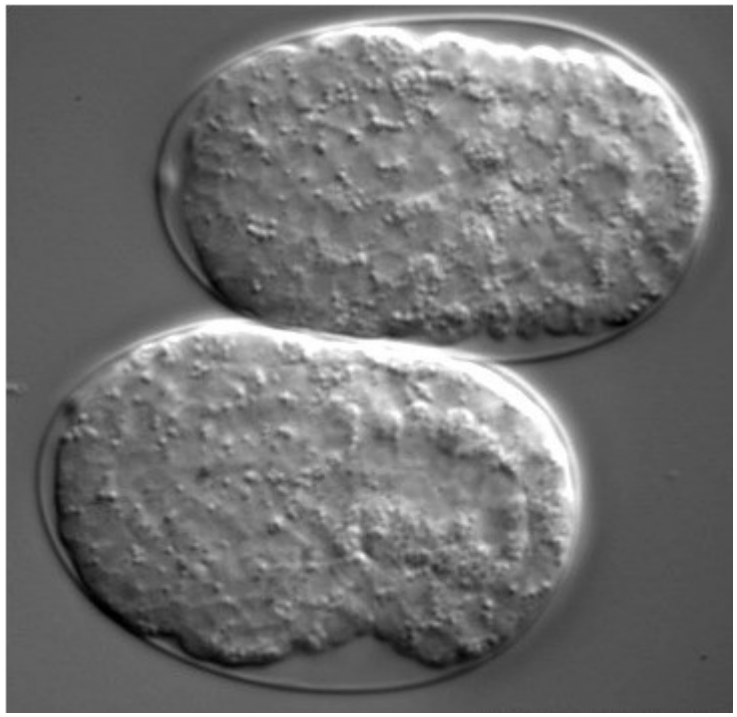
In mutanti del gene Ced-1 avviene l'apoptosi ma non la fagocitosi delle cellule morte (accumulo di cellule apoptotiche)



Ced-3, Ced-4 (in assenza di espressione di Ced-9)

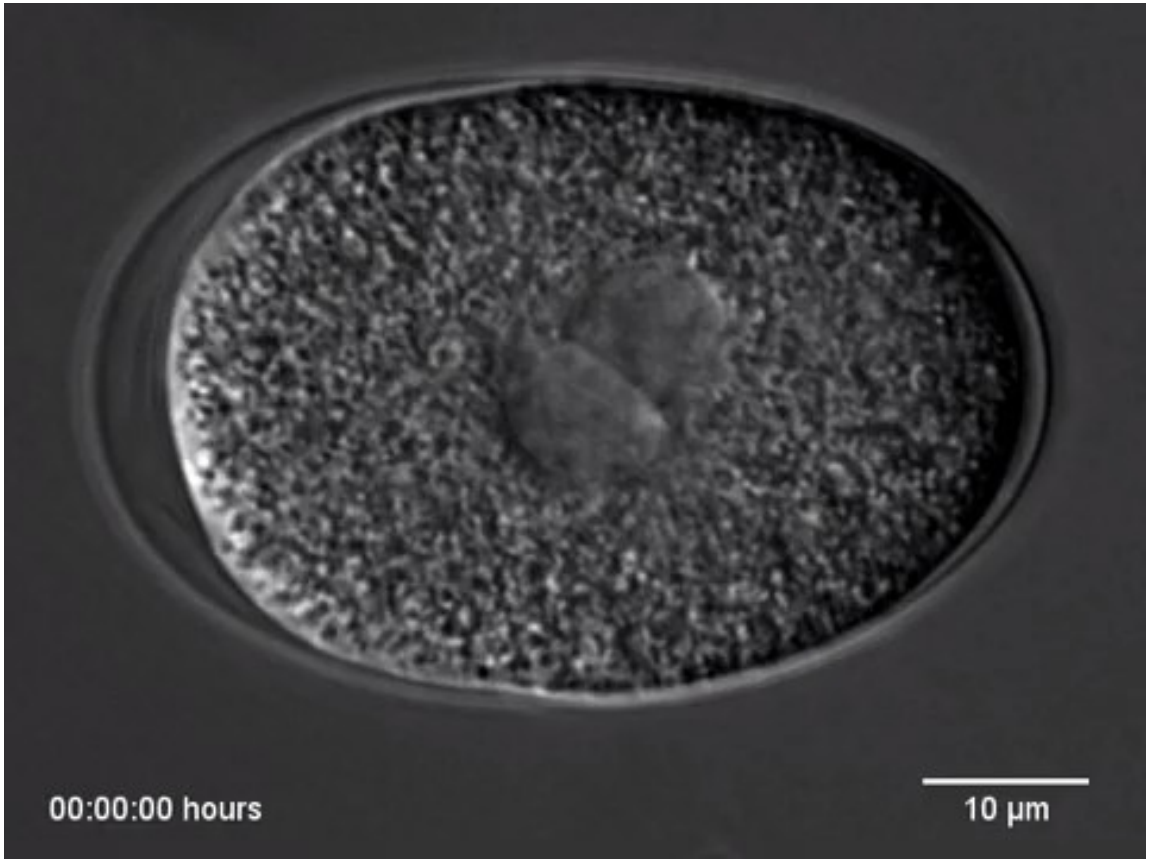
Ced-1 nella cellula fagocitante

C. Elegans mutants embryos

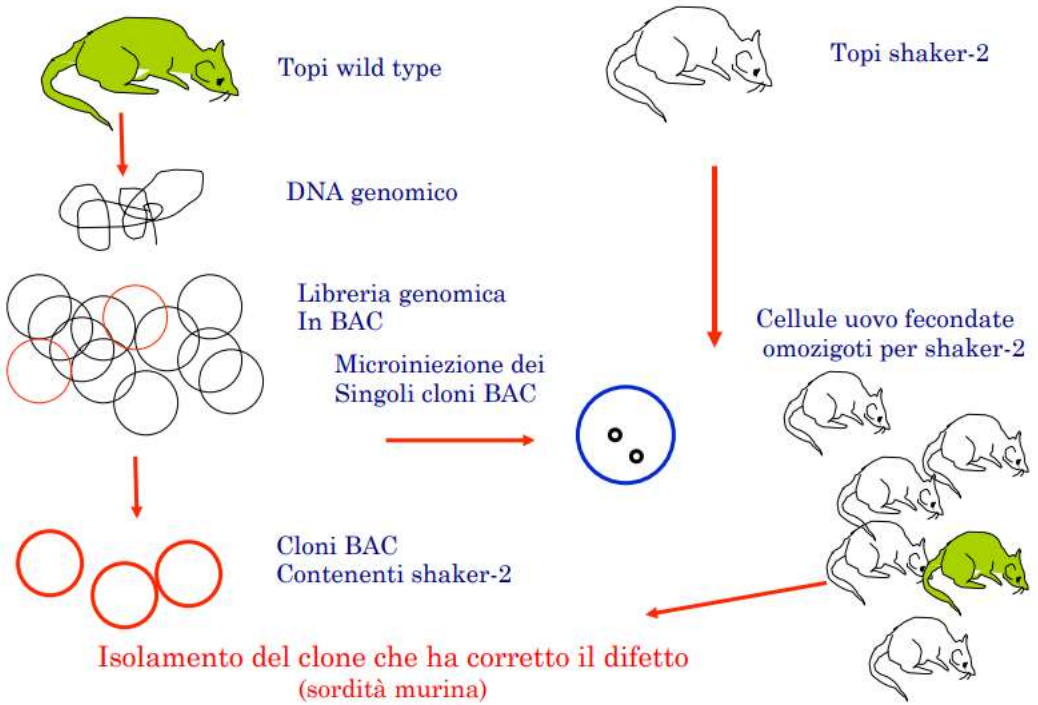


Mut. Ced1/Ced3

Mut. Ced1


00:00:00 hours
10 µm

Complementazione funzionale in topi trasgenici



La bioinformatica identifica un nuovo gene della miosina (myo 15)

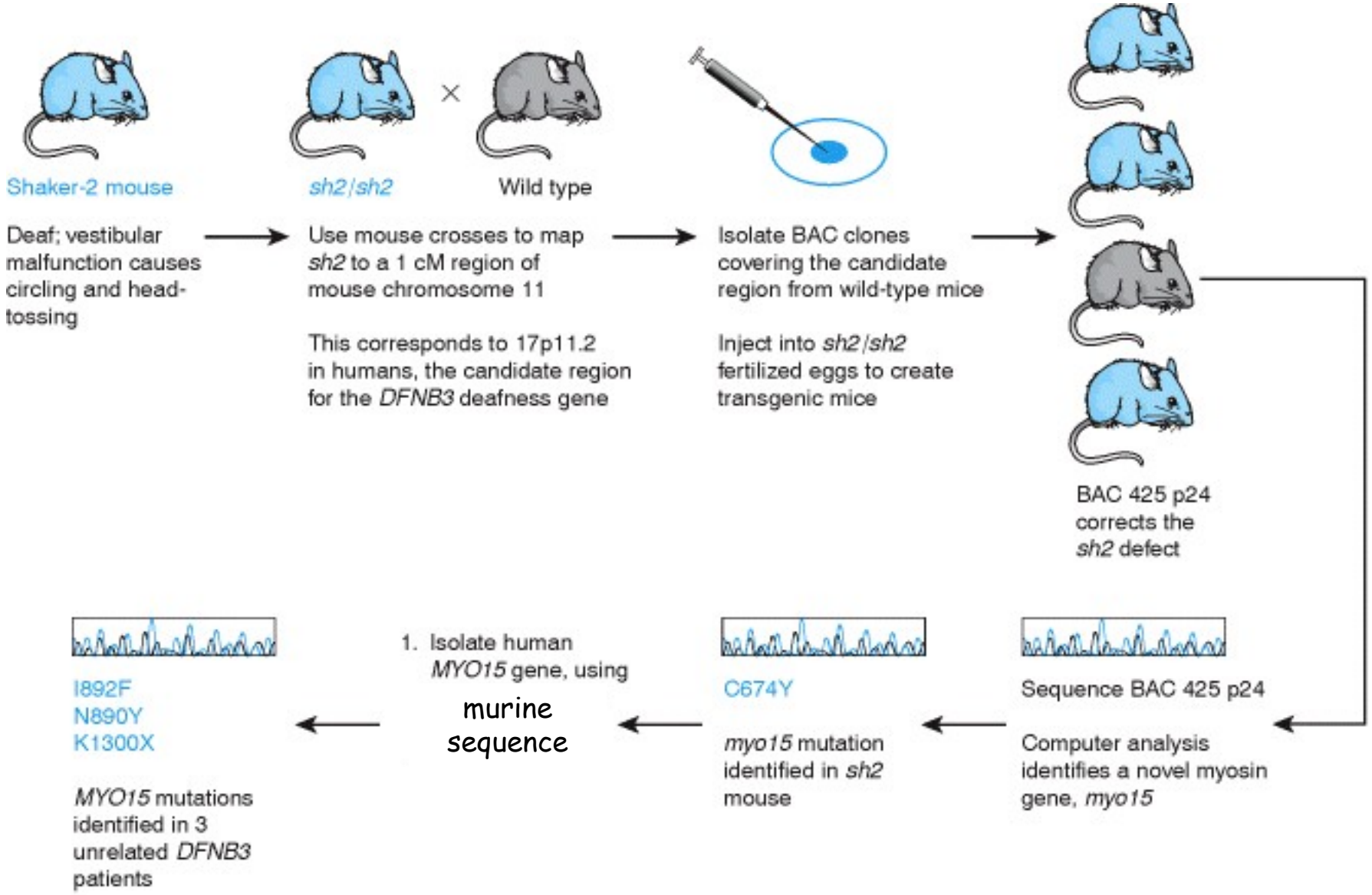
Una mutazione è stata identificata nel gene myo 15 del topo sh2

Sulla base della sequenza murina è stato identificato il gene umano Myo 15

I892F
N890Y
K1300X

Nel gene Myo 15di pazienti sordi sono state individuate mutazioni

Complementazione funzionale in topi trasgenici



Metodi per l'analisi funzionale dei geni:

-**inattivazione di un gene**

- **sovraespressione di un gene**

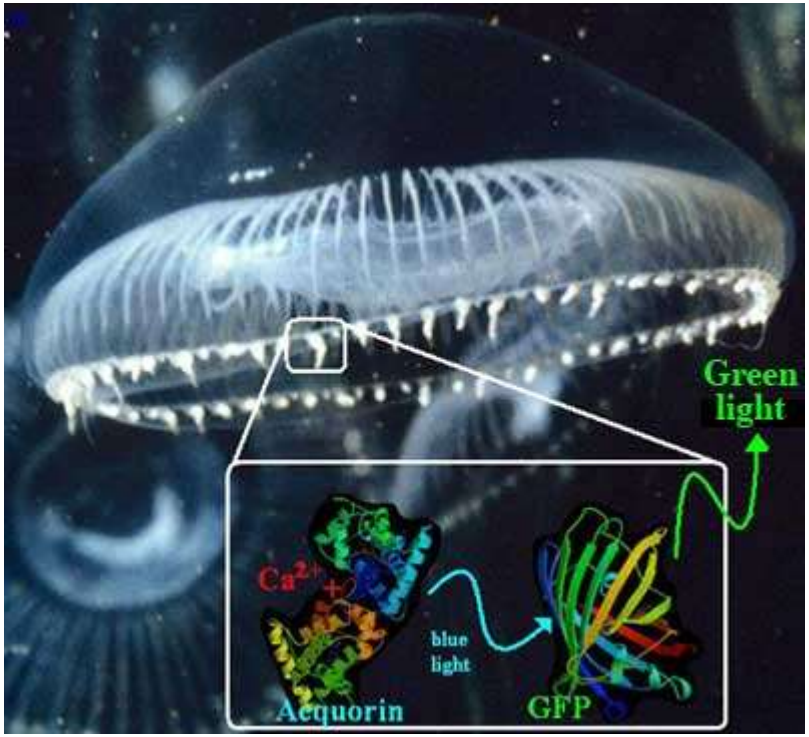
I geni reporter possono essere utilizzati per localizzare dove e quando i geni vengono espressi

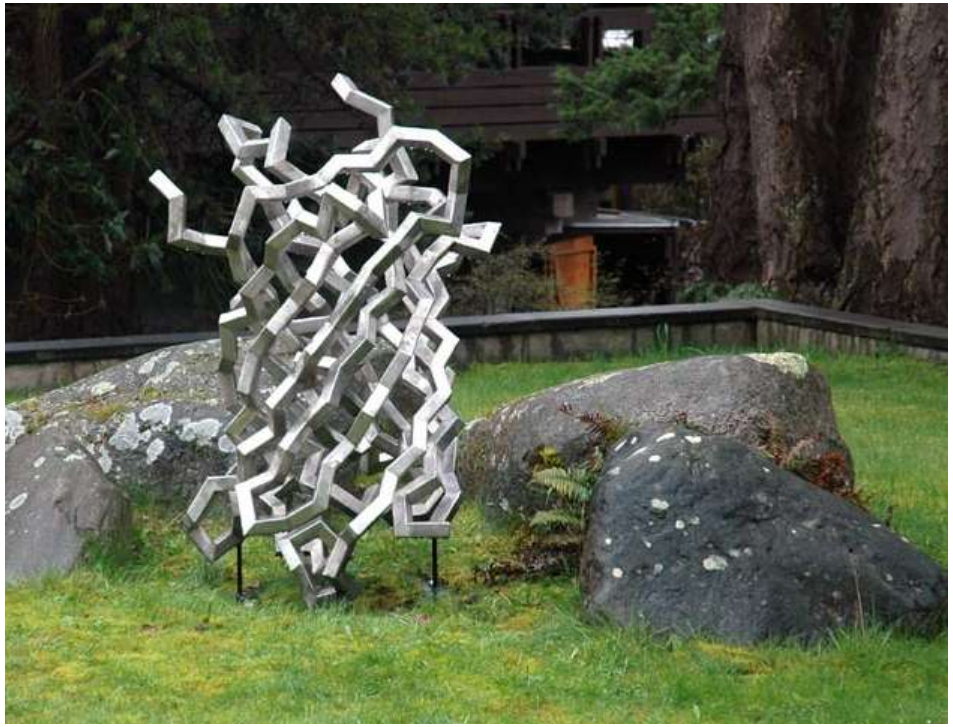
- Organi o tessuti specifici
- Particolari momenti dello sviluppo

- Geni reporter (promotore del gene più gene reporter) consentono di analizzare l'espressione del gene in modo semplice, idealmente tramite un esame visivo con cellule che diventano blu (saggio di analisi colorimetrica, reporter lacZ) o fluorescenti (analisi della fluorescenza, reporter Green fluorescent protein GFP).

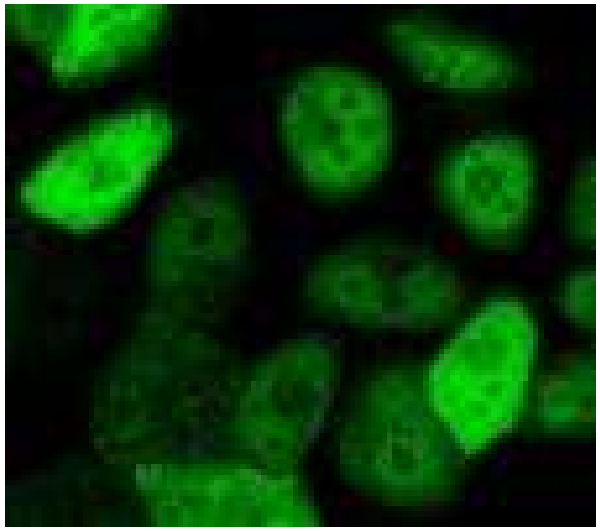
http://50annidna.scienze.unipd.it/DFTB/concept_34_ITA/con34problem.swf

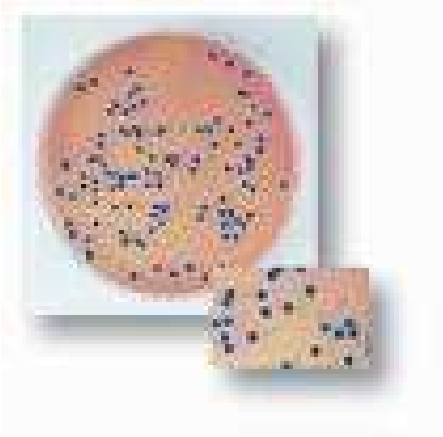
Osamu Shimomura, premio Nobel per la chimica nel 2008, scopre negli anni '60 la GFP





scultura in acciaio inossidabile presso i Friday Harbor Laboratories sull'Isola di San Juan (Washington, Stati Uniti), luogo della scoperta della GFP.





Il **gene batterico LacZ** espresso sotto la guida di promotori specifici